

# ECE531 Lecture 7: Bayesian Estimation and an Introduction to Non-Random Parameter Estimation

D. Richard Brown III

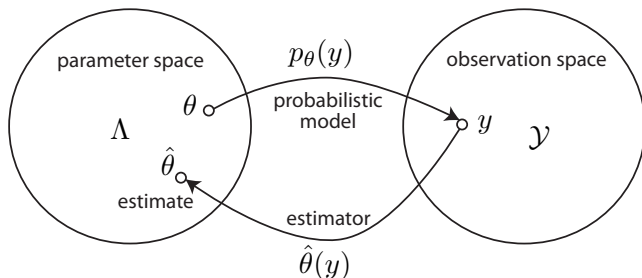
Worcester Polytechnic Institute

12-March-2009

# Introduction

- ▶ Hypothesis testing and detection: Make a choice between two (or several) discrete situations.
- ▶ Estimation: Determine as accurately as possible the actual value of the parameter(s) from the observation.
- ▶ Parameter (or point) estimation problems:
  - ▶ What prior information do we have about the parameter(s)?
  - ▶ What are our performance criteria (how much do we pay for bad estimates)?
- ▶ Two basic approaches:
  - ▶ Bayesian: The unknown parameter(s) have a known prior distribution.
  - ▶ Non-random: The unknown parameter(s) do not possess any known prior distribution.

# Mathematical Model for Estimation



Notation and terminology:

- ▶  $Y$  denotes the random observation with realizations  $y \in \mathcal{Y} \subseteq \mathbb{R}^n$ .
- ▶ The parameter space  $\Lambda$  is assumed to be a subset of  $\mathbb{R}^m$ .
- ▶ An **estimator** is a function mapping  $\mathcal{Y} \mapsto \Lambda$ .
- ▶ An **estimate** is a realization of the estimator corresponding to a particular observation  $y$ .
- ▶ The shorthand notation  $\hat{\theta}$  can mean the estimate or the estimator.

# Examples

Estimating the unknown frequency of a sinusoid in noise

$$Y_k = \cos(\theta k) + W_k \text{ for } k = 0, \dots, n-1 \text{ with } W_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\hat{\theta}(y) = \arg \max_{x \in [-\pi/2, \pi/2]} \left| \sum_{k=0}^{n-1} y_k e^{-jkx} \right|$$

Estimating the unknown interval of uniformly distributed observations

$$Y_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{U}(0, \theta) \text{ for } k = 0, \dots, n-1$$

$$\hat{\theta}(y) = \max\{y_0, \dots, y_{n-1}\}$$

Estimating the mean and variance of Gaussian distributed observations

$$Y_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2) \text{ for } k = 0, \dots, n-1$$

$$\hat{\mu}(y) = \frac{1}{N} \sum_{k=0}^{n-1} y_k \quad \hat{\sigma}^2(y) = \frac{1}{N} \sum_{k=0}^{n-1} (y_k - \hat{\mu})^2$$

# Cost Assignments and Conditional Risk

- ▶ Intuitively, an optimal estimator will find the “best” guess of the true parameter  $\theta$ .
- ▶ The solution to this problem depends on how we define “best” and how we penalize any deviation from “best”.
- ▶ Cost assignment:  $C_\theta(\hat{\theta}) : \Lambda \times \Lambda \mapsto \mathbb{R}$  is the cost of the parameter estimate  $\hat{\theta} \in \Lambda$  given the true parameter  $\theta \in \Lambda$ .
- ▶ Conditional risk of estimator  $\hat{\theta}(y)$  when the true parameter is  $\theta$ :

$$\begin{aligned} R_\theta(\hat{\theta}) &:= \mathbb{E} \left[ C_\theta(\hat{\theta}(Y)) \mid \theta \right] \\ &= \int_{\mathcal{Y}} C_\theta(\hat{\theta}(y)) p_\theta(y) dy \end{aligned}$$

# Some Common Cost Assignments

For any  $p \geq 1$  and  $x \in \mathbb{R}^m$ , the  $p$ -norm is

$$\|x\|_p := \left( \sum_{i=1}^m |x_i|^p \right)^{1/p}$$

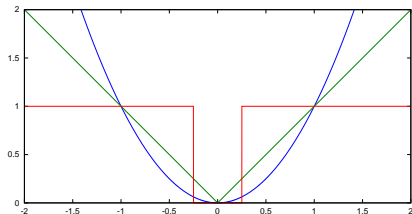
Let  $\epsilon := \hat{\theta}(y) - \theta$ . Many cost assignments can be written as  $C_\theta(\hat{\theta}) = C(\epsilon)$ .

- ▶ Squared error:  $C_\theta(\hat{\theta}) = \|\epsilon\|_2^2$ .
- ▶ Absolute error:  $C_\theta(\hat{\theta}) = \|\epsilon\|_1$ .
- ▶ Uniform error ("hit or miss"):

$$C_\theta(\hat{\theta}) = \begin{cases} 0 & \|\epsilon\|_\infty \leq \frac{\Delta}{2} \\ 1 & \text{otherwise} \end{cases}$$

where

$$\|x\|_\infty := \max\{|x_1|, \dots, |x_m|\}.$$



# The Bayesian Approach

We assume that the unknown parameter(s) are random with a known prior distribution  $\Theta \sim \pi(\theta)$ . The average/Bayes risk of estimator  $\hat{\theta}(y)$  is then

$$\begin{aligned}
 r(\hat{\theta}) &= \mathbb{E}[R_{\Theta}(\hat{\theta})] \\
 &= \int_{\Lambda} R_{\theta}(\hat{\theta})\pi(\theta) d\theta \\
 &= \int_{\Lambda} \int_{\mathcal{Y}} C_{\theta}(\hat{\theta}(y))p_{\theta}(y)\pi(\theta) dy d\theta \\
 &= \int_{\mathcal{Y}} \int_{\Lambda} C_{\theta}(\hat{\theta}(y))p_{\theta}(y)\pi(\theta) d\theta dy.
 \end{aligned}$$

This should look similar to composite Bayesian hypothesis testing where

$$r(\rho, \pi) = \int_{\mathcal{Y}} \rho^{\top}(y) \int_{\mathcal{X}} C(x)p_x(y)\pi(x) dx dy$$

where  $C(x) = [C_{0,x}, \dots, C_{M-1,x}]^{\top}$ . Our intuition here was to specify a decision rule that selected the index with “minimum commodity cost”.

# The Bayesian Approach

We can obtain additional intuition by rewriting the conditional density

$$p_x(y) := p_Y(y | X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)} = \frac{\pi(x | Y = y)p_Y(y)}{\pi(x)} = \frac{\pi_y(x)p(y)}{\pi(x)}$$

Hence, the Bayes risk for composite hypothesis testing can be written as

$$r(\rho, \pi) = \int_{\mathcal{Y}} \rho^\top(y) \left\{ \int_{\mathcal{X}} C(x) \pi_y(x) dx \right\} p(y) dy$$

When  $p(y) > 0$ , these problems are equivalent:

$$\min \int_{\mathcal{X}} C(x) \pi_y(x) dx p(y) \Leftrightarrow \min \int_{\mathcal{X}} C(x) \pi_y(x) dx$$

These problems are not equivalent, however, when  $p(y) = 0$ . Nevertheless, we can make arbitrary decisions on the set of observations  $\{y \in \mathcal{Y} : p(y) = 0\}$  without any effect on the Bayes risk.



# The Bayesian Approach

Back to Bayesian *estimation*. The same analysis can be used to write

$$\begin{aligned}
 r(\hat{\theta}) &= \int_{\mathcal{Y}} \int_{\Lambda} C_{\theta}(\hat{\theta}(y)) p_{\theta}(y) \pi(\theta) d\theta dy \\
 &= \int_{\mathcal{Y}} \underbrace{\int_{\Lambda} C_{\theta}(\hat{\theta}(y)) \pi_y(\theta) d\theta}_{\text{posterior cost of estimator } \hat{\theta}(y) \text{ when } Y=y} p(y) dy
 \end{aligned}$$

The Bayes estimate of the true parameter  $\theta$  can be found by minimizing this posterior cost for each  $y \in \mathcal{Y}$ . That is, we can fix  $y$  and solve the minimization problem

$$\begin{aligned}
 \hat{\theta}_{\text{opt}}(y) &= \arg \min_{g(\cdot)} \int_{\Lambda} C_{\theta}(g(y)) \pi_y(\theta) d\theta \\
 &= \arg \min_{g(\cdot)} \mathbb{E}[C_{\Theta}(g(y)) | Y = y]
 \end{aligned}$$

for each  $y \in \mathcal{Y}$ . The solution, of course, depends on our choice of  $C_{\theta}(\cdot)$ .

# Bayesian Estimation: Minimum Mean Squared Error

**Squared error cost assignment:**  $C_{\Theta}(g(y)) = \|g(y) - \Theta\|_2^2$ .

$$\begin{aligned}\hat{\theta}_{\text{mmse}}(y) &= \arg \min_{g(\cdot)} \mathbb{E}[\|g(y) - \Theta\|_2^2 | Y = y] \\ &= \arg \min_u u^T u - 2u^T \mathbb{E}[\Theta | Y = y] + \mathbb{E}[\Theta^T \Theta | Y = y]\end{aligned}$$

How do we solve this sort of problem? We can find the minimum by taking the gradient with respect to  $u$  and setting it equal to zero...

$$\nabla_u \left\{ u^T u - 2u^T \mathbb{E}[\Theta | Y = y] + \mathbb{E}[\Theta^T \Theta | Y = y] \right\} = 2u - 2\mathbb{E}[\Theta | Y = y]$$

hence

$$2u = \mathbb{E}[2\Theta | Y = y]$$

and we can conclude that  $\hat{\theta}_{\text{mmse}}(y) = \mathbb{E}[\Theta | Y = y]$ .

# Bayesian Estimation: Minimum Mean Absolute Error

**Absolute error cost assignment:**  $C_{\Theta}(g(y)) = \|g(y) - \Theta\|_1$ .

$$\begin{aligned}
 \hat{\theta}_{\text{mmae}}(y) &= \arg \min_{g(\cdot)} \mathbb{E}[\|g(y) - \Theta\|_1 \mid Y = y] \\
 &= \arg \min_u \mathbb{E} \left[ \sum_i |u_i - \Theta_i| \mid Y = y \right] \\
 &= \arg \min_u \sum_i \left\{ \int_{-\infty}^{u_i} (u_i - \theta_i) \pi_y(\theta_i) d\theta_i + \int_{u_i}^{\infty} (\theta_i - u_i) \pi_y(\theta_i) d\theta_i \right\} \\
 &= \arg \min_u \sum_i q(u_i, y)
 \end{aligned}$$

where  $\pi_y(\theta_i)$  is the posterior density of parameter  $\theta_i$  given the observation  $Y = y$ . The quantity  $q(u_i, y)$  is differentiable in  $u_i$ , so we can solve this easily using Leibnitz's Rule...

$$\frac{\partial}{\partial u_i} q(u_i, y) = \int_{-\infty}^{u_i} \pi_y(\theta_i) d\theta_i - \int_{u_i}^{\infty} \pi_y(\theta_i) d\theta_i$$

What value of  $u_i$  makes this go to zero? The median of the parameter  $\theta_i$ , conditioned on  $Y = y$ . Hence, we can conclude that  $\hat{\theta}_{\text{mmae}}(y) = \text{Median}[\Theta \mid Y = y]$ .

# Bayesian Estimation: Maximum A Posteriori Probability

Uniform cost assignment:

$$C_{\theta}(g) = \begin{cases} 0 & \|g(y) - \theta\|_{\infty} \leq \frac{\Delta}{2} \\ 1 & \text{otherwise} \end{cases}$$

We seek to find an estimator that minimizes the Bayes risk

$$\hat{\theta}_{\text{map}}(y) = \arg \min_{g(\cdot)} \mathbb{E}[C_{\theta}(g) | Y = y].$$

To find  $\hat{\theta}_{\text{map}}(y)$ , we can write

$$\begin{aligned} \hat{\theta}_{\text{map}}(y) &= \arg \min_{g(\cdot)} \int C_{\theta}(g) \pi_y(\theta) d\theta \\ &= \arg \min_{g(\cdot)} \left\{ 1 - \int_{g_1(y) - \Delta/2}^{g_1(y) + \Delta/2} \cdots \int_{g_m(y) - \Delta/2}^{g_m(y) + \Delta/2} \pi_y(\theta) d\theta \right\} \end{aligned}$$

# Bayesian Estimation: Maximum A Posteriori Probability

We can't take this much further without two additional assumptions: (A1) the posterior density  $\pi_y(\theta)$  is smooth and (A2)  $\Delta$  is small. Under these assumptions, we can write

$$\int_{g_1(y)-\Delta/2}^{g_1(y)+\Delta/2} \cdots \int_{g_m(y)-\Delta/2}^{g_m(y)+\Delta/2} \pi_y(\theta) d\theta \approx \Delta^m \pi_y(\theta)|_{\theta=g(y)}.$$

Hence we have

$$\hat{\theta}_{\text{map}}(y) = \arg \min_{g(\cdot)} \{1 - \Delta^m \pi_y(g)\}.$$

Since  $\Delta > 0$ , we can discard the  $\Delta^m$  term and use our usual tricks to write

$$\hat{\theta}_{\text{map}}(y) = \arg \max_{g(\cdot)} \pi_y(g).$$

Remarks:

- ▶  $\hat{\theta}_{\text{map}}(y) = \text{Mode}[\Theta | Y = y]$ .
- ▶ The solution is usually unique (but doesn't have to be).

# Bayesian Estimation: Summary of Common Approaches

- ▶ Bayesian MMSE, MMAE, and MAP estimators are distinguished only by the choice of cost assignment.

- ▶ Squared error cost assignment: MMSE

$$\hat{\theta}_{\text{mmse}}(y) = \text{E}[\Theta | Y = y].$$

- ▶ Absolute error cost assignment: MMAE

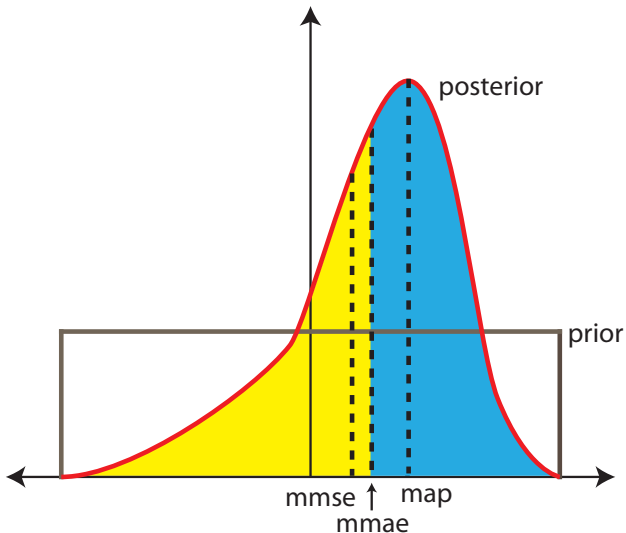
$$\hat{\theta}_{\text{mmae}}(y) = \text{Median}[\Theta | Y = y].$$

- ▶ Uniform cost assignment with smooth posterior and small  $\Delta$ : MAP

$$\hat{\theta}_{\text{map}}(y) = \text{Mode}[\Theta | Y = y].$$

- ▶ The observation  $Y = y$  converts the given prior distribution  $\pi(\theta)$  on the unknown parameter(s) to a posterior distribution  $\pi_y(\theta)$ .
- ▶ Each estimator is simply a feature of this posterior distribution.
- ▶ While the MAP solution is only approximate, it is widely used because MAP estimates are sometimes easier to compute than MMSE, MMAE, or other estimates.

## Bayesian Estimation: Summary of Common Approaches



# Bayesian Minimum Mean Squared Error

$$\text{MMSE} = \text{E} \left[ \|\Theta - \hat{\theta}_{\text{mmse}}(Y)\|_2^2 \right]$$

where the expectation is evaluated with respect to the joint pdf  $p_{Y,\Theta}(y, \theta)$ .

$$\begin{aligned} \text{MMSE} &= \int \int \|\theta - \text{E}[\Theta | Y = y]\|_2^2 p_{Y,\Theta}(y, \theta) dy d\theta \\ &= \int \int \|\theta - \text{E}[\Theta | Y = y]\|_2^2 \pi_y(\theta) d\theta p(y) dy \\ &= \int \int \sum_i (\theta_i - \text{E}[\Theta_i | Y = y])^2 \pi_y(\theta) d\theta p(y) dy \\ &= \int \sum_i \text{var}(\Theta_i | Y = y) p(y) dy \\ &= \int \text{trace} \{ \text{cov}(\Theta | Y = y) \} p(y) dy \end{aligned}$$

where  $\text{trace}(\cdot)$  is the sum of the diagonal elements of a matrix.



# Linear Gaussian Model

An important special case that covers many common estimation scenarios is the linear Gaussian signal model. In this model, the vector observation is given as

$$Y = H\Theta + W$$

where the observation  $Y \in \mathbb{R}^n$ , the “mixing matrix”  $H \in \mathbb{R}^{n \times m}$  is known, the unknown parameter vector  $\Theta \in \mathbb{R}^m$  is distributed as  $\mathcal{N}(\mu_\Theta, \Sigma_\Theta)$ , and the unknown noise vector  $W \in \mathbb{R}^n$  is distributed as  $\mathcal{N}(0, \Sigma_W)$ . Unless otherwise specified, we always assume the noise and the unknown parameters are independent of each other.

To specify a Bayesian estimator, we are going to need to compute posterior distribution  $\pi_y(\theta)$ . Note that we can't just write

$$\Theta = H^{-1}(Y - W)$$

since  $H$  may not be invertible (or even square). Hence, finding  $\pi_y(\theta)$  involves a little bit of work...

# Linear Gaussian Model

To develop an expression for the posterior distribution  $\pi_y(\theta)$ , we first note that  $\pi_y(\theta) = \frac{p_{Y,\Theta}(y,\theta)}{p_Y(y)}$ . To find the joint distribution  $p_{Y,\Theta}(y,\theta)$  let

$$Z = \begin{bmatrix} Y \\ \Theta \end{bmatrix} = \begin{bmatrix} H & I \\ I & 0 \end{bmatrix} \begin{bmatrix} \Theta \\ W \end{bmatrix}$$

Since  $\Theta$  and  $W$  are independent of each other and each is Gaussian, they are jointly Gaussian. Furthermore, since  $Z$  is a linear transformation of a jointly Gaussian random vector, it too is jointly Gaussian.

To fully characterize  $Z$ , we just need its mean and covariance:

$$\begin{aligned} \mu_Z &:= \mathbb{E}[Z] = \begin{bmatrix} H\mu_\Theta \\ \mu_\Theta \end{bmatrix} \\ \Sigma_Z &:= \text{cov}[Z] = \begin{bmatrix} H\Sigma_\Theta H^\top + \Sigma_W & H\Sigma_\Theta \\ \Sigma_\Theta H^\top & \Sigma_\Theta \end{bmatrix} \end{aligned}$$

# Linear Gaussian Model

$$\pi_y(\theta) = \frac{p_Z(z)}{p_Y(y)} = \frac{\frac{1}{(2\pi)^{(m+n)/2} |\Sigma_Z|^{1/2}} \exp \left\{ \frac{-(z-\mu_Z)^\top \Sigma_Z^{-1} (z-\mu_Z)}{2} \right\}}{\frac{1}{(2\pi)^{n/2} |\Sigma_Y|^{1/2}} \exp \left\{ \frac{-(y-\mu_Y)^\top \Sigma_Y^{-1} (y-\mu_Y)}{2} \right\}}$$

To simplify the terms outside of the exponentials, recall that

$$\Sigma_Z := \text{cov}[Z] = \begin{bmatrix} H\Sigma_\Theta H^\top + \Sigma_W & H\Sigma_\Theta \\ \Sigma_\Theta H^\top & \Sigma_\Theta \end{bmatrix} = \begin{bmatrix} \Sigma_Y & \Sigma_{Y,\Theta} \\ \Sigma_{\Theta,Y} & \Sigma_\Theta \end{bmatrix}$$

The determinant of a partitioned matrix  $P = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$  can be written as

$|P| = |A| \cdot |D - CA^{-1}B|$  if  $A$  is invertible. Covariance matrices are invertible, hence the terms outside the exponentials can be simplified to

$$\frac{\frac{1}{(2\pi)^{(m+n)/2} |\Sigma_Z|^{1/2}}}{\frac{1}{(2\pi)^{n/2} |\Sigma_Y|^{1/2}}} = \frac{1}{(2\pi)^{m/2} |\Sigma_\Theta - \Sigma_{\Theta,Y} \Sigma_Y^{-1} \Sigma_{Y,\Theta}|^{1/2}}$$

# Linear Gaussian Model

To simplify the terms inside the exponentials, a good approach is to use a matrix inversion formula for partitioned matrices ( $A$  must be invertible)

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}$$

and the matrix inversion lemma

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}.$$

Skipping all the algebraic details, we can write

$$\frac{\exp \left\{ \frac{-(z - \mu_Z)^\top \Sigma_Z^{-1} (z - \mu_Z)}{2} \right\}}{\exp \left\{ \frac{-(y - \mu_Y)^\top \Sigma_Y^{-1} (y - \mu_Y)}{2} \right\}} = \exp \left\{ \frac{-(\theta - \alpha(y))^\top \Sigma^{-1} (\theta - \alpha(y))}{2} \right\}$$

where  $\alpha(y) = \mu_\theta + \Sigma_{\theta,Y} \Sigma_Y^{-1} (y - \mu_Y)$  and  $\Sigma = \Sigma_\theta - \Sigma_{\theta,Y} \Sigma_Y^{-1} \Sigma_{Y,\theta}$ .

# Linear Gaussian Model

## Lemma

*In the linear Gaussian model, the parameter vector  $\Theta$  conditioned on the observation  $Y = y$  is jointly Gaussian distributed with*

$$\begin{aligned} \mathbb{E}[\Theta | Y = y] &= \mu_{\Theta} + \Sigma_{\Theta} H^{\top} \left( H \Sigma_{\Theta} H^{\top} + \Sigma_W \right)^{-1} (y - H \mu_{\Theta}) \\ \text{cov}[\Theta | Y = y] &= \Sigma_{\Theta} - \Sigma_{\Theta} H^{\top} \left( H \Sigma_{\Theta} H^{\top} + \Sigma_W \right)^{-1} H \Sigma_{\Theta} \end{aligned}$$

## Corollary

*In the linear Gaussian model*

$$\hat{\theta}_{mmse}(y) = \hat{\theta}_{mmae}(y) = \hat{\theta}_{map}(y)$$

# Linear Gaussian Model: Remarks

- ▶ All of the estimators are linear (actually affine) in the observation  $y$ .
- ▶ Note that  $\text{cov}[\Theta | Y = y]$  does not depend on  $y$ . Hence, the minimum Bayes risk for the quadratic cost function, i.e. the MMSE, is

$$\text{trace} \{ \text{cov}[\Theta | Y = y] \} = \text{trace} \{ \Sigma_{\Theta} \} - \text{trace} \left\{ \Sigma_{\Theta} H^{\top} \Sigma_Y^{-1} H \Sigma_{\Theta} \right\}.$$

## Example: Estimation of a Constant in White Noise

Suppose we have random observations given by

$$Y_k = \Theta + W_k \quad k = 0, \dots, n - 1$$

where  $W_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$  and  $\Theta \sim \mathcal{N}(\mu, v^2)$ . Let's derive the usual Bayesian estimators for the unknown parameter  $\Theta$ ...

# Example: Estimation of a Constant in White Noise

$$\hat{\theta}_{\text{mmse}}(y) = \text{E}[\Theta | Y = y] = \frac{\frac{v^2}{\sigma^2}n\bar{y} + \mu}{\frac{v^2}{\sigma^2}n + 1}$$

$$\text{MMSE} = \text{E}[\text{var}[\Theta | Y = y]] = \frac{v^2}{\frac{v^2}{\sigma^2}n + 1}$$

where  $\bar{y} := \frac{1}{n} \sum_{k=0}^{n-1} y_k$ . Remarks:

- ▶ When  $n = 0$ , the MMSE estimate  $\hat{\theta} = \mu$  and the MMSE is simply  $v^2$ .
- ▶ Note that  $v^2$  is a measure of the accuracy of our prior knowledge. If  $v^2$  is small, we know  $\Theta$  accurately without any observations.
- ▶ Note that MMSE is strictly decreasing in  $n$  as long as  $v > 0$ .
- ▶ The effect of the prior on  $\hat{\theta}_{\text{mmse}}$  also becomes less important with more samples. In the limit

$$\lim_{n \rightarrow \infty} \hat{\theta}_{\text{mmse}} = \bar{y} \quad \text{and} \quad \lim_{n \rightarrow \infty} \text{MMSE} = 0$$



## Example: Estimation of Signal Amplitude in Colored Noise

Suppose now that we have random observations given by

$$Y_k = \Theta s_k + W_k \quad k = 0, \dots, n-1$$

where  $s = [s_0, \dots, s_{n-1}]$  is known,  $W \sim \mathcal{N}(0, \Sigma)$ , and  $\Theta \sim \mathcal{N}(\mu, v^2)$ .  
Let's derive the usual Bayesian estimators for the unknown parameter  $\Theta$ ...

# Example: Deconvolution of a Gaussian Signal in Noise

Suppose we have a Gaussian distributed signal  $S_0, \dots, S_{L-1}$  transmitted through a multipath channel and corrupted by additive noise:

$$Y_k = \sum_{\ell=0}^{L-1} h_{k-\ell} S_{\ell} + W_k \quad k = 0, \dots, n-1$$

where the causal multipath channel impulse response  $h_0, h_1, \dots$  is known,  $W \sim \mathcal{N}(0, \Sigma)$ , and  $\Theta \sim \mathcal{N}(\mu, v^2)$ . This can be written in the linear Gaussian form by writing the convolution as a matrix-vector product:

$$\underbrace{\begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_{n-1} \end{bmatrix}}_Y = \underbrace{\begin{bmatrix} h_0 & 0 & \dots & 0 \\ h_1 & h_0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_{n-1} & h_{n-2} & \dots & h(n-L) \end{bmatrix}}_H \underbrace{\begin{bmatrix} S_0 \\ S_1 \\ \vdots \\ S_{L-1} \end{bmatrix}}_{\Theta} + \underbrace{\begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_{L-1} \end{bmatrix}}_W$$

# Conclusions: Bayesian Parameter Estimation

- ▶ Bayesian estimation assumes that the unknown parameters have a known prior pdf. The goal is to optimally estimate a particular realization of the random parameters by incorporating this knowledge.
- ▶ When applicable, the Bayesian approach can have better estimation accuracy than the types of estimators we will consider later (where no prior information is assumed).
- ▶ Bayesian MMSE, MMAE, and MAP estimators are the same only in special cases, e.g. the linear Gaussian model. See example IV.B.1 in your textbook for a case where all three estimators are distinct.
- ▶ Bayesian estimators have historically been controversial.
  - ▶ It is often difficult to specify and justify a meaningful prior pdf.
  - ▶ This motivates the study of “non-random” parameter estimation where no prior knowledge of the unknown parameter is assumed.

# Introduction to Non-Random Parameter Estimation

- ▶ In non-random parameter estimation problems, we can still compute the conditional risk of estimator  $\hat{\theta}(y)$  when the true parameter is  $\theta$ :

$$\begin{aligned} R_{\theta}(\hat{\theta}) &:= \mathbb{E}_{\theta} [C_{\theta}(\hat{\theta}(Y))] \\ &= \int_{\mathcal{Y}} C_{\theta}(\hat{\theta}(y)) p_Y(y; \theta) dy \end{aligned}$$

where  $\mathbb{E}_{\theta}$  means the expectation parameterized by  $\theta$ , i.e.  $\theta$  is fixed, and  $C_{\theta}(\hat{\theta}) : \Lambda \times \Lambda \mapsto \mathbb{R}$  is the cost of the parameter estimate  $\hat{\theta} \in \Lambda$  given the true parameter  $\theta \in \Lambda$ .

- ▶ We cannot, however, compute any sort of average risk

$$r(\hat{\theta}) = \mathbb{E}[R_{\Theta}(\hat{\theta})]$$

since we have no distribution on the random parameter  $\Theta$ .

Non-random parameter estimation does not mean that we know the parameter. It means that **we don't even have a prior for the parameter.**

# Basic Concepts of Non-Random Parameter Estimation

- ▶ We would like to find a “uniformly most powerful estimator”  $\hat{\theta}(y)$  that minimizes the conditional risk  $R_{\theta}(\hat{\theta})$  for all  $\theta \in \Lambda$ . Example:
  - ▶ Suppose  $\hat{\theta}(y) \equiv \theta_0 \in \Lambda$  (a constant) for all  $y \in \mathcal{Y}$ .
  - ▶ Then, for all of the cost functions we have considered,  $C_{\theta_0}(\hat{\theta}) = 0$ .
  - ▶ When the true parameter  $\theta = \theta_1 \neq \theta_0$ , however, the estimator  $\hat{\theta}(y) \equiv \theta_1$  outperforms  $\hat{\theta}(y) = \theta_0$  when  $\theta = \theta_1$ .
- ▶ It should be clear that a “uniformly most powerful estimator” is not going to exist in most cases of interest.
- ▶ Some options:
  1. We could restrict our attention to finding the sort of problems that do admit a “uniformly most powerful estimator”.
  2. We could try find “locally most powerful” estimators.
  3. We could assume a prior  $\pi(\theta)$ , e.g. perhaps some sort of least favorable prior, and solve the problem in the Bayes framework.
  4. We could keep the problem non-random but place restrictions on the class of estimators that we are willing to consider.

Important restricted class of estimators: **Unbiased estimators**.

## Option 4: Consider Only Unbiased Estimators

A reasonable restriction on the class of estimators that we are willing to consider is the class of **unbiased estimators**.

### Definition

An estimator  $\hat{\theta}(y)$  is unbiased if

$$\mathbb{E}_{\theta} [\hat{\theta}(Y)] = \theta$$

for all  $\theta \in \Lambda$ .

Remarks:

- ▶ This class excludes trivial estimators like  $\hat{\theta}(y) \equiv \theta_0$ .
- ▶ Under the squared-error cost assignment, estimators in this class

$$R_{\theta}(\hat{\theta}) = \mathbb{E}_{\theta} [\|\theta - \hat{\theta}(Y)\|_2^2] = \sum_i \mathbb{E}_{\theta_i} [(\hat{\theta}_i(Y) - \theta_i)^2] = \sum_i \text{var}_{\theta_i} [\hat{\theta}_i(Y)]$$

- ▶ Optimal approach: find unbiased estimators with minimum variance.

# Minimum Variance Unbiased Estimators

## Definition

A minimum-variance unbiased estimator  $\hat{\theta}_{\text{mvu}}(y)$  is an unbiased estimator satisfying

$$\hat{\theta}_{\text{mvu}}(y) = \arg \min_{\hat{\theta} \in \Omega} R_{\theta}(\hat{\theta})$$

for all  $\theta \in \Lambda$  where  $\Omega$  is the set of all unbiased estimators.

Remarks:

- ▶ Finding an MVU estimator is still a multi-objective optimization problem.
- ▶ MVU estimators do not always exist.
- ▶ The class of problems in which MVU estimators do exist, however, is much larger than that of “uniformly most powerful” estimators.